Large Scale Data Processing

J. HAYS

PPTAP SOFTWARE AND COMPUTING MINI-WORKSHOP

19TH JULY 2021

Queen Mary
University of London

Challenges: Outline

“Large Scale”

What are the challenges of running systems “at scale”?

“Data”

What are the challenges around data handling?

“Processing”

What are the challenges around processing and compute?

Summary

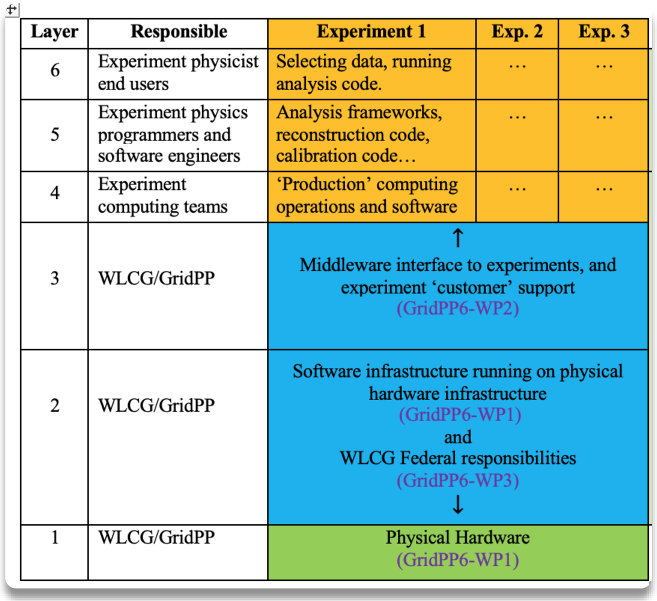
Challenges in running at scale

- Not controversial to say that processing, data handling, and networking needs will be increasing for Particle Physics (and beyond) in the coming years
  - Systems running at scale bring their own challenges
  - Organisation, coordination and management
  - Distributed systems
  - Interoperability
  - Access methods and identify management
  - On-boarding and user support

Queen Mary
University of London

Challenges in running at scale

- Support models and funding channels need to match the needs
  - DevOps, ResOps, and RSE support
  - Funding must be targeted in the right way
    - Likely means a mix of approaches
    - Needs to be planned with appropriate horizons – (eg not year-to-year)
- Interoperability between communities and systems
  - Access methods and Identity Management
  - User support and on-boarding
- Resource management
  - Fair access and usage, representation across areas,



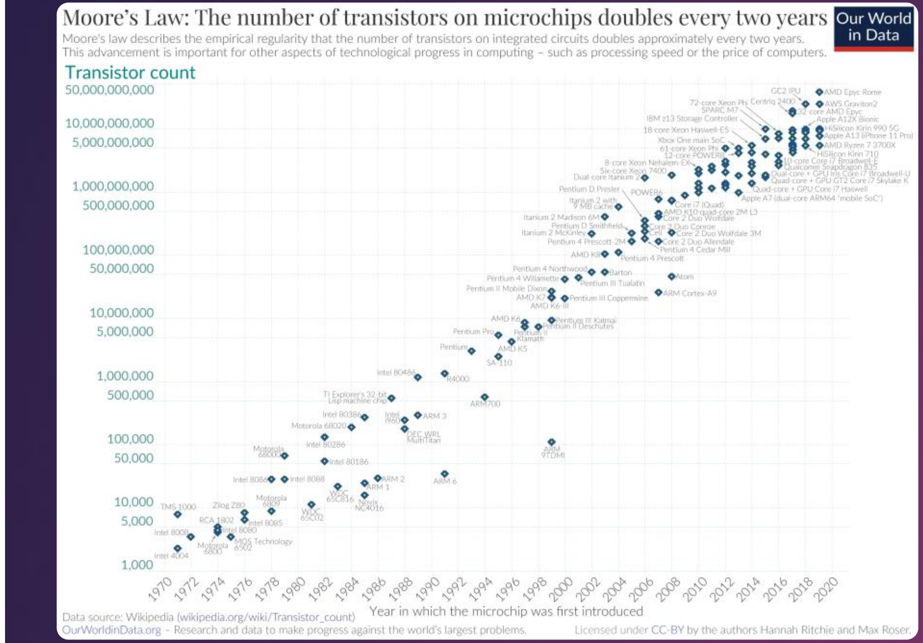
Layer	Responsible	Experiment 1	Exp. 2	Exp. 3
6	Experiment physicist end users	Selecting data, running analysis code.	…	…
5	Experiment physics programmers and software engineers	Analysis frameworks, reconstruction code, calibration code…	…	…
4	Experiment computing teams	'Production' computing operations and software	…	…
3	WLCG/GridPP	↑ Middleware interface to experiments, and experiment 'customer' support (GridPP6-WP2)		
2	WLCG/GridPP	Software infrastructure running on physical hardware infrastructure (GridPP6-WP1) and WLCG Federal responsibilities (GridPP6-WP3) ↓		
1	WLCG/GridPP	Physical Hardware (GridPP6-WP1)		

Challenges with data

- Large data volumes
  - Data handling and management
    - Moving data around requires networks to support it
    - Support data handling policies – replication etc
  - Data integrity
    - Custodial data services
  - Data storage
    - Hot data, cold data, archival data

See the talk by Alastair…

Challenges with Compute



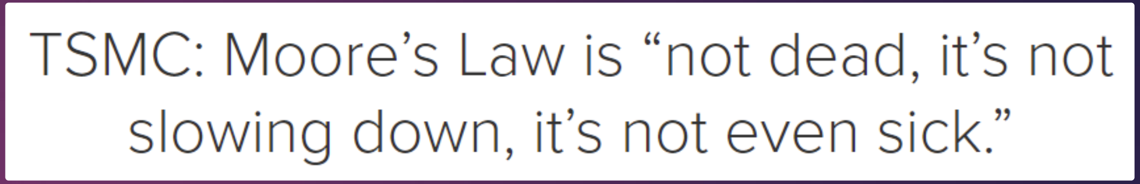
Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

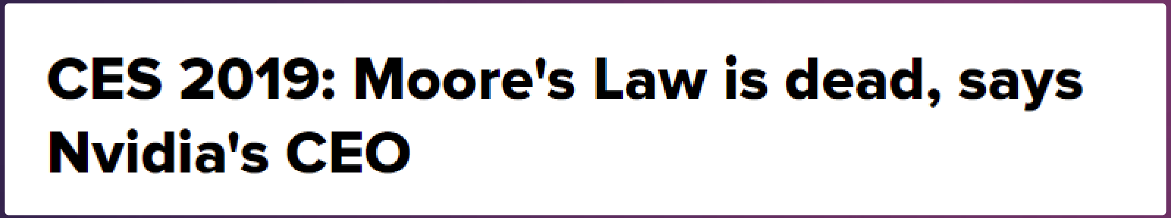
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

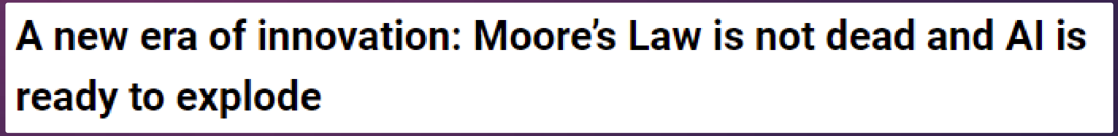
Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.



TSMC: Moore's Law is "not dead, it's not slowing down, it's not even sick."



CES 2019: Moore's Law is dead, says Nvidia's CEO



A new era of innovation: Moore's Law is not dead and AI is ready to explode

Challenges with compute

- Either way – dead or alive
- Substantially increasing demand for compute resources
  - New(ish) technologies and architectures
    - GPU, FPGA, IPU, APU, … ?
    - Software to take advantage of it
    - Evaluation – benchmarking and evidence-based decision making
    - Accounting
  - Heterogeneous systems and workloads
    - Software infrastructure to join it up

Queen Mary
University of London

Needs trained people to succeed but also needs careful management to make sure benefits are shared/captured across multiple areas

The future is heterogeneous!

Challenges with compute

- Either way – dead or alive…

- Substantially increasing demand for compute resources
  - Where to put it?
  - Power, cooling and environmental concerns
    - Sustainability
  - Shared versus dedicated infrastructure
  - Funding and allocation

Queen Mary
University of London

Summary

- Significant challenges
  - Hardware
  - Software
  - People
  - Environment
  - Cost

New technologies will need new ways of working

Blurring existing boundaries

Investment in people, infrastructure and organisation needed to reap the benefits

Funding must be targeted in the right way

Queen Mary
University of London